PEPTIDE DERIVATIZATION FOR ENHANCING PROTEIN IDENTIFICATION BY MASS SPECTROMETRY

RELATED APPLICATIONS

This application claims priority under 35 USC §119(e) to U.S. Provisional Application Serial No. 60/523,643, filed November 20, 2003, the disclosure of which is incorporated herein by reference.

US GOVERNMENT RIGHTS

This invention was made with United States Government support under Grant No. 5R01 GM61336-4, awarded by the National Institutes of Health. The United States Government has certain rights in the invention.

BACKGROUND

The investigation of biological systems by mass spectrometry has rapidly evolved in recent years due to a wealth of advancements in both instrument design and bioinformatics. While the development of this field is ongoing, a number of technologies now exist that greatly facilitate the characterization of complex biological mixtures. An important component of this type of work is the ability to confidently identify proteins in an expeditious manner. The two most common approaches used to achieve this goal are tandem mass spectrometry and MALDI mass mapping. In either type of experiment peptides generated by proteolysis are analyzed following some form of chromatographic or electrophoretic separation. Subsequently, proteins are assigned by comparing the mass spectrometric data to theoretical sequences in a database.

In a typical proteomic experiment thousands of unknowns may be interpreted using automated search routines such as SEQUEST or MASCOT. These algorithms compare MS/MS spectra to the hypothetical fragment ion masses of database sequences and calculate a score for each match that quantifies the likelihood of an assignment. This general approach to protein identification has been successfully utilized in many different types of experiments. However, database matching does possess limitations. Since candidate sequences for assignments are generated using precursor ion masses these algorithms will often mishandle database errors, genetic mutations, and modifications that occur either during sample-handling or post-translationally.

Considering the numerous types of peptide modifications in database matching is often not practical since the complexity of a search can increase exponentially leading to prohibitively large databases that increases false-positive assignments and search times. Since the mapping of post-translational modification (PTM) sites is often critically important for deciphering the function of proteins, this represents a serious drawback of the present techniques. Furthermore, organisms without a sequenced genome cannot be studied using database matching techniques. In light of these limitations there is a need for methods that extract information from spectra independent of databases.

A number of different *de novo* sequencing approaches have been developed in recent years to help achieve this goal. The most straightforward approach to *de novo* sequencing is to make interpretations using the mass differentials between consecutive peaks of the same ion series. However, this seemingly simple task represents a significant challenge for a number of reasons. Peptides do not typically yield a contiguous series of ions that would enable complete sequencing. Furthermore, the discernment of N- and C-terminal fragment ions is not straightforward since both types are commonly formed by most activation methods. Mistakes in sequencing will result from this ambiguity if peaks from different ion series' (e.g. b- and y-ions) are used together in calculating mass differentials. Independent of these problems, sequencing errors may also occur as a result of the similar masses of lysine (128.0950 u) and glutamine (128.0586 u) residues, as well as the isobaric leucine and isoleucine (113.0841 u each) residues.

Accordingly, it is highly desirable to have a peptide derivatization strategy that utilizes labels that lead to more predictable fragmentation patterns and/or impart a mass code that allows N- and C-terminal fragment ions to be distinguished. One aspect of the present invention is directed to a *de novo* sequencing method that utilizes both guanidination of lysine residues in conjunction with amidination of the N-termini of the peptides to be analyzed by mass spectrometry. This approach facilitates identification of N- and C-terminal fragment ions by labeling N-termini with amidine moieties that differ by a methylene group (i.e. 14 u). In addition, the conversion of lysine residues to homoarginines prevents amidination of the side-chain ϵ-amino groups. These simple and efficient reactions are inexpensive and can be completed rapidly with minimal side-reactions.

SUMMARY OF VARIOUS EMBODIMENTS OF THE INVENTION

In accordance with one illustrative embodiment of the present invention there is provided a covalent derivatization strategy for *de novo* peptide sequencing. In particular, a method of the present invention facilitates the identification of proteins and their post-translational modifications via *de novo* interpretation of peptide sequences in tandem mass spectrometry. In an illustrative first step, lysine residues are blocked by, for example, guanidination, and subsequently, peptide N-termini are selectively labeled with, for example, either acetamidine or propionamidine groups. This labeling scheme enables distinction between N- and C-terminal fragment ions when MS/MS spectra of labeled peptide ions are compared. N-terminal fragment ions (a-, b-, and c-type) appear with mass differentials of 14 Da divided by the charge, while C-terminal fragment ions (x-,y-, and z-type) are isobaric. Accordingly, one aspect of the present invention provides a method of identifying a protein or peptide by searching a genomic database utilizing sequence information that is derived by directly interpreting mass spectral data.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates tandem MS spectra of guanidinated/amidinated peptides and their unmodified counterparts.

Fig. 2A & 2B illustrate the Q-TOF tandem mass spectra of the unlabeled peptide (Fig. 2A) and labeled peptide (Fig. 2B) $[M+2H]^{2+}$ EFTPPVQAAYQK (SEQ ID NO: 3) precursor ion.

Fig. 3A & 3B illustrate the Q-TOF tandem mass spectra of the acetamidinated peptide (Fig. 3A) and the propionamidinated peptide (Fig. 3B) FFVAPFPEVFGK (SEQ ID NO: 4) precursor ions.

Fig. 4A & 4B illustrate the Q-TOF tandem mass spectra of the acetamidinated peptide (Fig. 4A) and the propionamidinated peptide (Fig. 4B) YLGYLEQLLR (SEQ ID NO: 6) precursor ion.

Fig. 5A & 5B illustrate the Q-TOF tandem mass spectra of the acetamidinated peptide (Fig. 5A) and the propionamidinated peptide (Fig. 5B) LLVVYPW (SEQ ID NO: 7) precursor ion.

Fig. 6A & 6B illustrate the Q-TOF tandem mass spectra of the acetamidinated peptide (Fig. 6A) and the propionamidinated peptide (Fig. 6B) $[M+2H]^{2+}$ VPQLEIVPN(pS)AEER phosphopeptide (SEQ ID NO: 9) precursor ion.

DETAILED DESCRIPTIONS OF ILLUSTRATIVE EMBODIMENTS

Definitions

In describing and claiming the invention, the following terminology will be used in accordance with the definitions set forth below.

The term "peptide" as used herein encompasses a sequence of 2 or more amino acids joined to each other by peptide bonds. Peptides may contain amino acids other than the 20 gene-encoded amino acids, and includes amino acid sequences modified either by natural processes, such as post-translational processing, or by chemical modification techniques which are well known in the art. Such modifications are well described in basic texts, as well as in the research literature. Modifications can occur anywhere in a peptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. It will be appreciated that the same type of modification may be present in the same or varying degrees at several sites in a given peptide. Also, a given peptide may contain many types of modifications. See, for instance, Proteins-Structure and Molecular Properties, 2nd Ed., T. E. Creighton, W. H. Freeman and Company, New York, 1993 and Wold, F., Posttranslational Protein Modifications: Perspectives and Prospects, pgs. 1-12 in Posttranslational Covalent Modification of Proteins, B. C. Johnson, Ed., Academic Press, New York, 1983; Seifter et al., "Analysis for protein modifications and nonprotein cofactors", Methods in Enzymol. 182:626-646 (1990) and Rattan et al., "Protein Synthesis: Posttranslational Modifications and Aging", Ann NY Acad Sci 663:48-62 (1992).

As used herein the term "subjecting to mass spec analysis" includes not only the steps of determining the mass spectra of the precursor ions of the peptide, but also the steps of interpreting the mass spectral data (i.e. using mass spacings between adjacent peaks of same type to identify residues) to directly derive sequence information regarding the peptide.

Embodiments

The present invention is directed to a novel strategy for enhancing the mass spectrometric analysis of peptides. In accordance with one embodiment, the peptides to be analyzed are derivatized by labeling the N-terminus with amidine groups. More particularly, the N-termini of peptides to be analyzed are labeled with an S-methyl thioimidine. In one embodiment the N-termini of the peptide is labeled with an acetamidine group and/or a propionamidine group. Applicants have discovered that

labeling the N-terminus with amidine groups promotes specific fragmentation pathways that facilitate *de novo* sequencing by providing sequence information that is often absent. More particularly, N-termini derivatizations with amidine groups promotes the cleavage of the N-terminal peptide bond. Therefore, abundant y_{n-1} and b_1 ions are typically observed in MS/MS spectra. These fragment ions provide sequence information that is typically unavailable from unmodified peptides (i.e. unmodified peptides will often not yield contiguous fragment-ion series) and allow reliable interpretation of N-terminal residues. Applicants have found that b_1 ions can serve as internal calibrants to achieve mass accuracies of less than 10 ppm. This attribute should facilitate *de novo* sequencing and could also be used to further constrain database searching strategies. This information can be used to dramatically improve mass accuracy.

In accordance with one embodiment of the present invention peptide N-termini are selectively labeled with two amidine groups that differ in molecular weight. For example in one embodiment peptide N-termini are selectively labeled with either acetamidine or propionamidine groups. These two groups are structurally homologous, differing only by a single methylene group. This labeling scheme enables distinction between N- and C-terminal fragment ions when MS/MS spectra of labeled peptide ions are compared. N-terminal fragment ions (a-, b-, and c-type) appear with mass differentials of 14 Da divided by the charge, while C-terminal fragment ions (x-,y-, and z-type) are isobaric. In one embodiment the peptide N-termini are selectively labeled with either acetamidine or propionamidine groups by *S*-methyl thioacetimidate and *S*-methyl thiopropionimidate, respectively.

In one embodiment the original composition comprising the peptide is divided into two groups prior to labeling the N-terminus of the peptide. Typically, the original peptide containing composition is divided into separate and distinct first and second pools of peptides in a manner whereby the content of the two pools of peptides is very similar, if not identical. In this embodiment the N-terminus of the peptides contained in the first pool are labeled with an S-methyl thioimidine that differs in molecular weight from the S-methyl thioimidine used to label the second pool of peptides. In accordance with one embodiment the N-termini of the first pool of peptides are labeled with an acetamidine group, and the N-termini of the second pool of peptides are labeled with a propionamidine group. In one embodiment the N-termini of the first pool of peptides are labeled utilizing S-methyl thioacetimidate, and the N-termini of the

second pool of peptides are labeled utilizing S-methyl thiopropionimidate. The two pools of amidinated samples are combined and the combined sample is analyzed by mass spectrometer.

In another embodiment the original composition comprising the peptide is divided into two groups prior to labeling the N-terminus of the peptide. Typically, the original peptide containing composition is divided into separate and distinct first and second pools of peptides in a manner whereby the content of the two pools of peptides is very similar, if not identical. In this embodiment the N-terminus of the peptides contained in the first pool are labeled with an amidine group, and the N-termini of the peptides of said second pool of peptides is labeled with the same amidine group, but the amidine group of the second pool of labeled peptides comprises an isotopic substituted group. **In accordance with one embodiment the isotopically substitution comprises one or more hydrogens substituted with deuterium. In another embodiment the isotopically substituted group is one or more carbons (C^{12}) substituted with C^{13}. Typically, only a single atom is substituted for an isotope. Advantageously, using labels that differ based on an isotope substitution will allow the two amidine tagged peptide counterparts co-elute, thus allowing the mass spectrometer to analyze the two pairs at the same time. In one embodiment the amidine group is selected from the group consisting of acetamidine, propionamidine, butyramidine and pentylamidine, including straight chained as well as branched derivatives.**

This approach of labeling the N-terminus of the peptide to be analyzed with two different amidine moieties facilitates identification of N- and C-terminal fragment ions. More particularly, in the embodiment wherein the respective pools of peptides are labeled utilizing S-methyl thioacetimidate and S-methyl thiopropionimidate, the N-terminal fragments will differ by a methylene group (i.e. 14 u). Accordingly, this labeling allows the N- and C-terminal ~~fragments~~ to be easily distinguished, and facilitates the interpretation of MS/MS data. The distinction of N- and C-terminal fragments is not possible without labeling. Moreover, peptide sequence coverage is generally improved by the above described labeling scheme, since N-termini derivatizations with amidine groups promotes cleavage of the N-terminal peptide bond, yielding abundant y_{n-1} and b_1 ions in MS/MS spectra.

For peptides that contain lysine residues, the peptides are typically first modified to prevent the amino functionality of the lysine group from reacting with the S-

methyl thioimidine. In accordance with one embodiment the lysine residues of the peptide are blocked through a guanidination reaction. In one embodiment the lysine residues are converted to homoarginine residues by guanidination with S-methylisothiourea or O-methylisourea. In one embodiment the lysine residues are converted to homoarginine residues by guanidination with S-methylisothiourea. It should be appreciated that lysine can be blocked by derivatizations other than guanidination as long as they don't react with the N-terminal amine. Subsequent to the guanidination of the peptides, the guanidinated peptides are separated into two separate and distinct pools and their respective N-termini are labeled with amidine moieties. In one embodiment the respective pools of peptides are labeled utilizing S-methyl thioacetimidate and S-methyl thiopropionimidate. The two pools of amidinated samples are combined and the combined sample is analyzed by mass spectrometer.

One advantage of the presently described labeling strategy is that labeled lysine residues no longer have a similar mass to glutamine, and can be more definitively assigned. The guanidination of all lysine residues shifts their masses from 128 to 170 Da, thus eliminating the overlap that exists between lysine (128.095 Da) and glutamine (128.059 Da). The small mass difference of 0.036 Da between unmodified lysine and glutamine is hard for most mass spectrometers (with the exception of FTICR) to distinguish and is a common problem that confuses peptide sequencing. The present invention eliminates this complication.

Furthermore, the N-termini labeling strategy of the present invention will also lead to an increase in MALDI ionization yields of many peptides since the highly basic labels promote protonation. In particular, the effect of the labels on MALDI ionization yields is particularly impressive when a peptide does not possess basic amino acid residues (e.g. lysine, arginine, and histidine). Peptides without such residues typically exhibit poor ionization yields. In fact, some of these peptides are only detectable after being amidinated via the present invention since this derivatization introduces a strongly basic group at the N-terminus.

It should also be appreciated that the method of fragmentation described herein is not limited to low-energy CID in an ion-trap. For example, CID experiments with doubly labeled peptides using a quadrupole time-of-flight mass spectrometer have been successfully performed. In addition, it should be possible to use this labeling approach in combination with any other form of activation (e.g. high-energy CID, ECD,

SID, photodissociation, IRMPD, BIRD) and mass analyzer (e.g. FTICR, TOF). The identification of proteins from *de novo* sequences is also possible by homology searching (i.e. BLAST).

The present invention also encompasses the protein and peptide derivatives produced in accordance with the present invention. More particularly, one embodiment of the present invention is directed to a set of modified proteins and peptides, and in one embodiment a set of modified tryptic peptide. The set of modified peptides comprises a first pool of peptides wherein N-termini of the peptides are labeled with an acetamidine group and a second pool of peptides wherein N-termini of the peptides are labeled with a propionamidine group, wherein the two pools of peptides are separate and distinct. Typically the two pools of peptides will be substantially the same but for the N-terminal labels added to the peptides. In another embodiment of the present invention the two pools are combined to provide a composition comprising a mixture of peptides having their N-termini labeled with an acetamidine group and peptides having their N-termini labeled with a propionamidine group. In one embodiment the mixture contain substantially equivalent amounts of peptides having their N-termini with an acetamidine group and peptides having their N-termini labeled with a propionamidine group.

In one embodiment the lysine residues of the peptides comprising the first and second pools of peptides are converted to homoarginines. In a further embodiment the N-termini of the peptides of said first pool of peptides are labeled with methyl thioacetimidate, and the N-termini of the peptides of said second pool of peptides are labeled with methyl thiopropionimidate. The two pools of amidinated samples are combined and the combined sample is analyzed by mass spectrometer.

One aspect of the present invention is directed to an improved method of identifying proteins or peptide by utilizing tandem mass spectrometry in conjunction with genomic database searching. This is made possible by the applicants' labeling procedure that enhances the ability to obtain amino acid sequence data from a given peptide or protein. However, when genomic database searching is utilized, such a search will be used in a substantially different way from what is currently done by standard commercial programs such as *Sequest* and *Mascot*. For example, rather than searching a database with a precursor mass to generate candidate sequences (the calculated fragments from which are all matched against experimental measured masses), database searches will be

performed using sequence information that is derived by directly interpreting the mass spectral data of the present invention. By utilizing an appropriate algorithm to generate sequence information from measured fragment masses, the derivatization approach described herein will facilitate the identification of proteins or peptides. This searching method will be faster and far more selective than searching databases using masses alone. The result will be more reliable protein identifications performed in much shorter time.

Since the described labeling approach resolves the ambiguities associated with distinguishing N- and C-terminal fragment ions, peptide sequences can be derived directly from the data. After identifying peptides using sequence-based database matching, the measured precursor and fragment masses will be compared with those predicted for the identified peptides. Performing this mass matching *after* the sequence matching will provide another level of analysis that will only increase the reliability of the method and will enable the identification of post-translational modifications. For cases in which the genome of an organism has not been sequenced or for some reason is unknown, protein identification with *de novo* sequencing can be performed by homology searching. This involves comparing observed sequences with those from other organisms. Accordingly, the approach described herein is especially useful when studying organisms whose genomes have not been sequenced. Database searching methods such as Sequest and Mascot *require* that the genome of an organism is known.

In accordance with one embodiment of the present invention, a method of identifying a protein or peptide by searching a genomic database utilizing sequence information that is derived by directly interpreting mass spectral data is provided. The method of obtaining at least a partial amino acid sequence of an unknown protein or peptide comprises the steps of blocking the lysine residues of a peptide to be analyzed through guanidination, labeling the N-termini of the peptide with a compound selected from the group consisting of an acetamidine group and a propionamidine group, and subjecting the labeled peptides to mass spectral analysis. In one embodiment the original peptide containing composition is divided into a first and second pool of peptides, and the N-termini of the peptides of the first pool of peptides are labeled with an acetamidine group, and the N-termini of the peptides of the second pool of peptides is labeled with a propionamidine group. The two pools of amidinated samples are then combined and the combined sample is analyzed by mass spectrometer. In accordance with one embodiment, the guanidination step is performed with S-methylisothiourea, the first pool

of peptides is labeled utilizing S-methyl thioacetimidate, and the second pool of peptides is labeled utilizing S-methyl thiopropionimidate. In one embodiment the dual labeled peptide is subjected to tandem MS/MS mass spectral analysis. With respect to the present invention, it should be appreciated that the mass-coded peptide N-termini facilitates the interpretation of MS/MS data since N- and C-terminal fragments can be easily distinguished. Since peptides are mass coded at their N-termini, this technique is a global approach to protein identifications.

In accordance with one embodiment a method of identifying a protein or peptide comprises the steps of blocking the lysine residues of the peptide with guanidination, labeling the N-termini of a portion of the peptide with an acetamidine group and labeling the N-termini of the remaining portion of the peptide with a propionamidine group, subjecting the labeled peptides to mass spectral analysis and determining at least a partial amino acid sequence of the protein or peptide. Typically the protein is subjected to proteolysis, such as trypic digestion prior to the step of blocking the lysine residues. The interpreted amino acid sequence is then used in database searches to identify proteins that contain such a sequence. For example the interpreted amino acid sequence can submitted to a Blast search of the NCBI reference sequence database. Although Blast searching provides the capability to match homologous sequences, in one embodiment the search can be limited to exact matches to reduce the number of hits. In cases where the amino acid sequence alone is not sufficient to provide an unambiguous match, the precursor mass can also be employed as a constraint. In this embodiment, a sequence match is only considered a proper assignment if the interpreted sequence was contained within a predicted peptide that was consistent with the observed precursor ion mass. As described in Example 3 (Table 2) the use of this simple constraint eliminated false positive matches and uniquely identified model proteins.

Although using the precursor masses and interpreted sequence is anticipated to be sufficient for identifying most proteins, it may be necessary to further constrain some searches. This would be especially important if only a short segment of a peptide (i.e. < 5 residues) was interpretable. As demonstrated in Table 2 most interpreted sequences begin with the N-terminal residue. It is clear that these sequences contain the N-terminus since the analysis begins with the b_1 and y_{n-1} fragment ions that are produced by amidinated peptides. In cases such as these, it would be possible to further limit random matches by requiring that the N-terminus of candidate peptides is contained in the

interpreted sequence. Another strategy to further refine assignments would be to use smaller fragments of interpretable sequences in addition to the contiguous sequences shown here. In all of the interpretations shown in Table 2 the longest contiguous sequence that was interpretable was matched against a database. However, it is often possible to identify shorter portions of a peptide sequence as well. Incorporation of this additional sequence information could be useful, especially in cases where a long contiguous sequence is not interpretable.

The protein identification method of present invention reduces the occurrence of false-positives since sequence information (i.e. sequence of residues *and* their modifications) will be derived from the data prior to database comparisons. This approach also leads to much quicker data interpretation since far fewer candidate sequences need to be considered. A typical proteomic data set requires several days for interpretation. This problem is greatly exacerbated when all of the possible combinations of post-translational modification are considered. Using conventional techniques, assignment of post-translational modifications is very difficult using database matching approaches since an algorithm must consider a prohibitive number of combinations of modifications. This problem leads to more false-positive assignments and far longer search times. For this reason, it is common practice to consider only unmodified peptides. This failure to interpret post-translational modifications is *a great loss* since these often are critical to biological pathways. The double-labeling approach of the present invention facilitates identification of post-translational modifications. By interpreting data in a *de novo* manner, post-translational modifications are always considered.

While the disclosure has been illustrated and described in detail in the foregoing description, such illustration and description is to be considered as exemplary and not restrictive in character, it being understood that only the preferred embodiments have been shown and described and that all changes and modifications that come within the spirit of the disclosure are desired to be protected. The following examples are intended only to further illustrate the invention and are not intended to limit the scope of the subject matter which is defined by the claims.

EXAMPLE 1

Mass Spectrometry Analysis of Modified vs. Unmodified Peptides

Methods and Procedures

Synthesis of S-methyl thioacetimidate

11 g of thioacetamide were dissolved in 1 L of anhydrous diethyl ether at ambient temperature with stirring. To this solution, 8.8 mL of iodomethane were added and the mixture was allowed to stand at room temperature for 14 h. A light yellow precipitate was collected by vacuum filtration. The powder was stored over desiccant at ambient temperature and was not further purified.

Synthesis of S-methyl thiopropionimidate

1.8 g of thiopropionamide were dissolved in 100 mL of 99.5% pure acetone. This solution was placed in a 60 °C water bath and 3.8 mL of iodomethane were added. The reaction mixture was allowed to stand for 1 h at the bath temperature. Dark yellow crystals were collected after completely evaporating the solvent in a vacuum chamber at ambient temperature. The crystals were stored at ambient temperature over desiccant and were not further purified.

Labeling of tryptic peptides

Lysine residues were guanidinated using *S*-methyl isothiourea. A stock solution of this reagent was prepared in 6% (m/v) ammonium hydroxide to a concentration of 1 mol/L. The stock solution was mixed 1:1 with tryptic peptide samples and incubated for 1 h at a temperature of 65 °C. Amidination reactions were performed on aliquots of guanidinated tryptic digests. The guanidinated samples were prepared for amidination by simply evaporating the ammonium hydroxide. Acetamidination was performed by mixing equal volumes of a digest aliquot and a 43.4 g/L solution of S-methylthioacetimidate that was dissolved in 250 mM tris-(hydroxymethyl)aminomethane. Likewise, propionamidination was performed by mixing equal volumes of the digest and a 46.2 g/L stock solution of *S*-methyl thiopropionimidate that was dissolved in the same buffer. Reaction mixtures were allowed to stand at ambient temperature for 1 h prior to addition of TFA to a concentration of 1.0% (v/v). Amidination reactions were performed in a fume hood.

Results

Doubly labeled and unmodified tryptic digests of several model proteins were analyzed by LC-MS/MS using an ion trap mass spectrometer (LCQ-Deca XP, Thermo Finnigan). As an example, a comparison of the MS/MS spectra of acetamidinated, propionamidinated, and unmodified VDPVNFK (SEQ ID NO: 1) and VLGAFSDGLAHLDNLK (SEQ ID NO: 2) are presented in Fig. 1. Both peptides were electrosprayed and VDPVNFK (SEQ ID NO: 1) was fragmented as a doubly charged ion while the latter was a triply charged ion. By using the "mass-coded" fragmentation data of the labeled peptides each could be extensively sequenced. In the case of VDPVNFK (SEQ ID NO: 1) a complete y-ion series was observed (y_1-y_6). Similarly, (y_3-y_{15}) was observed from VLGAFSDGLAHLDNLK(SEQ ID NO: 2). Using the methodology of the present invention y-ions were easily identifiable because their masses are the same regardless of the N-terminal label. Once the ion-types are distinguished one can deduce the amino acid sequence by simply calculating the mass differentials between adjacent y-ions. For example, the mass difference of 115 Da between y_6 and y_5 from VDPVNFK (SEQ ID NO: 1) is interpreted as an aspartic acid residue (D). Likewise the N-terminal residue, valine, is easily identified by calculating the mass differential between y_6 and the precursor ion. The observation of intense y_{n-1} ions (such as y_6 and y_{15} in Fig. 1) is another novel advantage of this technique. This fragment ion is not typically observed from unmodified peptides. Therefore, it would not be possible to determine the N-terminal residue in most cases without amidination of the N-terminus. Although very few b-ions are typically observed, it is still important to identify these as N-terminal fragments in order to avoid incorrect interpretations of sequences. Without the mass-labeling of b-ions it is possible for the mass difference between an observed y-ion and a b-ion to match a given residue mass leading to misinterpretation.

EXAMPLE 2

Fragmentation of Amidinated Peptide Ions

Materials

The proteins cytochrome c (horse), hemoglobin (human), serum albumin (bovine), carbonic anhydrase II (bovine), pyruvate kinase (rabbit), and TPCK-treated trypsin (bovine) were obtained from Sigma (St. Louis, MO). α-cyano-4-hydroxycinnamic

acid (CHCA) and tris-(hydroxymethyl)aminomethane were also purchased from Sigma. Anhydrous diethyl ether, thioacetamide, and ammonium bicarbonate were supplied by Fisher (Fair Lawn, NJ). Iodomethane, formic acid, and 2,5-dihydroxybenzoic acid (2,5-DHB) were purchased from Aldrich (Milwaukee, WI). Acetonitrile and trifluoroacetic acid (TFA) were obtained from EM Science (Gibbstown, NJ).

Labeling Tryptic Peptides

S-methyl thioacetimidate was synthesized as described in Example 1. Tryptic digests were prepared by combining model proteins with TPCK-treated trypsin (1:100 protein to trypsin molar ratio) in 25 mM ammonium bicarbonate and incubating this mixture for 12 h at a temperature of 37°C. These mixtures were acetamidinated by mixing equal volumes of a digest aliquot and a 43.4 g/L solution of S-methylthioacetimidate that was dissolved in 250 mM tris-(hydroxymethyl)aminomethane. These reactions were incubated at ambient temperature for 1 h prior to addition of TFA to a concentration of 2.0% (vol/vol). The synthesis of S-methyl thioacetimidate and peptide labeling reactions were performed in a fume hood.

Fragmentation of Multiply Protonated Peptides

Both unmodified and acetamidinated tryptic peptides were analyzed in LC-MS/MS experiments. Samples were injected onto a 1 mm i.d. C-18 reversed phase column (Grace Vydac, Hesperia, CA) and eluted with a linear gradient of organic modifier. The gradient was delivered at a flow rate of 50 μL/min ranging from 95% Solvent A, 5% Solvent B (A = 0.1% formic acid in water and B = 0.1% formic acid in acetonitrile) to 60% A, 40% B over 60 min. The effluent was split such that 90% of the flow was directed to waste while 10% was delivered to the ESI source. An ion trap mass spectrometer (LCQ-Deca XP Plus, Thermo Finnigan, San Jose, CA) was used for all experiments involving electrospray. MS/MS experiments were performed using a data dependent precursor ion selection strategy. Therefore, the most abundant ion in a full MS scan was selected for CID in the subsequent scan event. Full MS scans were acquired using automatic gain control (AGC) and an *m/z* range of 400–1700. Once isolated, precursor ions were activated by applying a narrowband (±1 u) resonant RF excitation waveform for 30 ms. The activation energy was normalized by adjusting the amplitude of the resonance excitation RF voltage to compensate for the *m/z*-dependent fragmentation of precursor ions. This voltage is directly proportional to precursor m/z

and the available range of voltages is established by setting an arbitrarily defined "normalized collision energy" value. In all experiments involving multiply charged peptides the normalized collision energy was set to a value of 35%. Also, an activation Q of 0.25 was applied in these studies.

Analysis of Singly Protonated Peptides

MALDI mass spectrometry was employed in the study of singly charged peptides using both ion trap (Thermo Finnigan LCQ Deca XP Plus with a Mass Tech atmospheric pressure source) and time of flight (Bruker Reflex III, Bremen, Germany) mass analyzers. MALDI spots were prepared in these experiments using CHCA matrix. This compound was dissolved in a solvent composed of 50% acetonitrile (vol/vol) and 0.1% TFA(vol/vol) in water to a concentration of 10 g/L. Peptide samples were combined with the matrix solution in a1:9 volumetric ratio and 1 μL of this mixture was deposited onto a probe. Ion trap mass spectra were acquired using both MS and MS/MS modes, but with a modification to the method used in ESI experiments. Full MS spectra of tryptic digests were acquired over an *m/z* range of 315–2000 without using automatic gain control. Instead, the ion injection time was maintained at 300 ms. This injection time is much higher than that typically employed in electrospray analyses with AGC, and was chosen to be compatible with the low repetition rate (10 Hz) of the AP/MALDI ion source. The CID of these peptides was performed using a normalized collision energy of 50%, an activation Q of 0.25, an activation time of 30 ms and, unless otherwise noted, wideband activation. The latter enabled excitation of ions having masses up to 20 u less than the precursor. This allowed us to further break down large fragment ions that were abundantly generated by the loss of NH_3 from precursor ions. A normalized collision energy of 50%, rather than the 35% employed in the electrospray experiments, helped to compensate for the loss of sensitivity for product ions that is common when wideband activation is used.

Reflectron MALDI-TOF mass spectra of an acetamidinated tryptic digest of hemoglobin were acquired using either 2,5-DHB or CHCA as the matrices. MALDI spots were prepared with 2,5-DHB by mixing 1 μL of matrix solution with 0.5 μ L of the labeled hemoglobin digest on probe. 2,5-DHB was dissolved to a concentration of 40 g/L in 20% acetonitrile (vol/vol) and 0.1% TFA (vol/vol) in water to make the matrix solution. MALDI spots were prepared with CHCA as above except only 0.7 μL of the

matrix/analyte solution was deposited on probe. In all MALDI spot preparations the acetamidinated digest mixtures were used without any purification prior to mixing with matrix solutions.

Results

Charge State Distribution Shifts

To investigate the phenomenon, that MALDI ion yields of amidine-labeled peptides exceeded those of their unmodified counterparts, the charge state distributions of electrospray ionization mass spectra of acetamidinated and unmodified tryptic peptides was compared. Mass spectra of these peptides were acquired between MS/MS scans during an LC-MS analysis as described in the experimental section. In all, 26 unmodified and acetamidinated peptides were compared. The peptides used in this study were derived from the tryptic digests of several model proteins including hemoglobin, cytochrome c, carbonic anhydrase, and serum albumin. In an attempt to gauge the relative propensity for acetamidinated and unmodified peptides to form multiply charged ions the charge state distributions of these peptides were compared by calculating an average charge state for each case. These values were determined by weighting the contributions of particular charge states based on their relative intensities. Therefore these comparisons do not reflect the total ion yields of each peptide.

Amidination was determined to increase the average charge state value in almost every case. We have considered the possibility that the different eluting conditions of amidinated and unmodified peptides could play a role in the observed charge state distributions since previous studies have indicated that higher concentrations of acetonitrile generally lead to increased average charge states in ESI and amidinated peptides elute at approximately 1% (vol/vol) higher acetonitrile in reversed phase LC. To probe this issue, peptides from identical electrospray conditions were analyzed. Labeled and unlabeled peptides from a tryptic digest of hemoglobin were purified by reversed phase LC, collected into the same solution and simultaneously electrosprayed by direct infusion. The results of this analysis were in excellent agreement with the data from the original LC experiments. Accordingly, the amidine labels rather than the solvent composition led to the observed increased protonation of peptides.

Fragmentation of Electrosprayed Amidine Labeled Peptides

The effect of amidination on the fragmentation of electrosprayed peptides was considered next. For this purpose, both unmodified and acetamidinated tryptic peptides of several proteins (pyruvate kinase, hemoglobin, carbonic anhydrase II, and serum albumin) were analyzed in LC-MS/MS experiments. Only the precursor ions that were at least doubly charged were considered. In total, the tandem mass spectra of 29 peptides were observed for which a direct comparison between the unmodified and acetamidinated forms could be made. Furthermore, MS/MS spectra of a total of 51 acetamidinated and 41 unmodified peptides from these digests were acquired. The fact that not every peptide was paired was most often the result of multiple components co-eluting. Therefore, an MS/MS spectrum could not be acquired in some cases because there was not enough time to perform CID on every component. This problem was exacerbated in our experiments since to attain more reliable results we repeated the MS/MS scans of a precursor ion five times before it was placed on an exclusion list and other ions could be analyzed. It would have been possible to obtain more labeled and unmodified pairs for direct comparison by repeating the analysis of these samples and focusing on unpaired peptides by placing their precursor *m/z* values on a priority list. However, since a significant number of peptide pairs were already detected, this was not deemed to be necessary.

In each comparison the precursor ions differed only by the presence of amidine groups at their N-termini and lysine residues. It is apparent from the data obtained that the addition of amidine labels induces significant changes in fragmentation. The most striking of these is the strongly increased production of y_{n-1} fragment ions from cleavage of the N-terminal residue. These y_{n-1} fragment ions are the most intense peaks in each of the acetamidinated MS/MS spectra. In contrast, y_{n-1} ions were often not even observed from unmodified peptides. Interestingly, the charge state of the y_{n-1} ion varied from one peptide to another.

Of the 51 acetamidinated peptides that were analyzed a y_{n-1} fragment ion was observed for every case. Furthermore, this ion was the base peak of its spectrum in 32 out of 51 cases (63%). By comparison, 18 out of 41 (44%) of the unmodified peptides also yielded y_{n-1} fragment ions, and in only one case (2%) was it the most intense peak in its spectrum. Despite the increased efficiency of N-terminal residue cleavages, the number of other peaks in MS/MS spectra of labeled peptides were comparable to that

observed with unmodified peptides. The fact that other sequence ions are observed from amidinated peptides should facilitate protein identifications. If one exploits the information that is available from the enhanced fragmentation of the N-terminal peptide bond, it is possible to increase the confidence of peptide assignments from database searches. The identification of the N-terminal residue provides a database searching constraint.

The application of this constraint was simulated using the translated genome of *Caulobacter crescentus* and a database analysis program (PRODIGIES) that was written in house [Nierman et al., C. M. Complete Genome Sequence of Caulobacter crescentus. Proc. Natl. Acad. Sci. U.S.A. 2001, 98, 4136–4141; and Karty et al., Defining Absolute Confidence Limits in the Identification of Caulobacter Proteins by Peptide Mass Mapping. J. Proteome Res. 2002, 1,325–335.]. When candidate sequences are limited by both the mass of a precursor ion and the identity of the N-terminal residue, the number of candidate sequences is reduced by approximately one order of magnitude. This simplification will reduce the occurrence of false positive sequence matches, thus generally improving the confidence of protein assignments. Furthermore, with fewer viable candidates, the amount of time required for database searches should decrease.

Enhanced Neutral Loss of NH3 in MALDI of Amidinated Peptides

One might expect different results for singly charged peptide ions since the charge might be sequestered on an amidino group and consequently be less mobile than in multiply protonated species. Thus we investigated the fragmentation of amidinated $[M + H]^{1+}$ ions using MALDI mass spectrometry. Data for acetamidinated tryptic digests of hemoglobin revealed that the loss of NH_3 occurs quite readily with CHCA matrix but not to a great extent with 2,5-DHB. The promotion of this type of fragmentation must be related to the addition of amidine labels since these results were not observed from this unmodified tryptic digest analyzed using the same conditions. Furthermore, the loss of NH3 appears to occur independent of lysine amidination. Lastly, a mass spectrum of this tryptic digest was acquired using AP/MALDI and an ion trap mass spectrometer with CHCA as the matrix. Compared with the CHCA/TOF data, fewer ions lose NH_3. This reduction of NH_3 loss may be explained by a combination of effects: Collisional cooling is faster when the ions are formed at atmospheric pressure, and the ions are also cooled by He buffer gas once injected into the ion trap.

CID of Singly Charged Acetamidinated Peptides

Using an AP/MALDI source and a quadrupole ion trap mass spectrometer we have investigated the CID of singly protonated, acetamidinated peptides. As in the electrospray study described above, tryptic peptides from the digests of several model proteins were employed in this work and the goal here was to compare the fragmentation of amidinated and unmodified peptides. The data were processed using the average spectra of 50 MS/MS scans, since averaged mass spectra more reliably reflect fragmentation tendencies than do single spectra. The y_{n-1} fragment ions were typically among the most abundant in the labeled proteins, while this fragment was very weak or not detected from the unmodified peptides. The enhancement of this dissociation pathway is similar to that observed with electrosprayed doubly and triply charged peptide ions. However, there are also some important charge-dependent differences. Most striking is the predominance of b-NH_3 (b*) fragment ions from singly charged amidinated precursor ions. In some cases a contiguous series of these ions was observed while the unmodified version of these peptides only produced fewer b-type fragments. The tendency of amidinated peptides to produce contiguous b*-ion series may be very useful in *de novo* sequencing experiments. Unmodified peptides often do not yield such complete and easily interpretable information. While this unique type of fragmentation may facilitate *de novo* sequencing, it is important to note that not every peptide generates b*-ions. Interestingly, the presence of proline residues in peptides often suppresses the formation of b*-ions. Many researchers have identified the formation of y-ions via cleavage on the N-terminal side of proline residues as a very efficient fragmentation pathway. It has been proposed that this pathway is generally favored because the proline's amide group is more efficiently protonated than others. Perhaps the high fragmentation efficiency from these sites precludes the formation of b*-ions.

Wideband Versus Narrowband Excitation

In the AP/MALDI experiments just discussed the use of wideband activation was important for minimizing the intensities of otherwise dominant $[M + H - NH_3]^+$ product ions. Limited fragmentation from amidinated peptides was commonly observed with narrowband excitation. In contrast, the use of wideband activation provided much more complete fragmentation. Therefore, product ions resulting from neutral losses of small groups such as NH_3 could be further activated to produce informative sequence ions. Unlike the experiment with narrow band activation, these

data were not dominated by NH_3 loss from the precursor ion. Wideband activation of peptide ions typically generated primarily b*- and y-type fragment ions. Furthermore, the $[M + H - NH_3]^+$ ion of each precursor was not detected when using wideband activation. The removal of this product is advantageous in protein identification experiments since it does not convey sequence-specific information. Interestingly, the types of sequence ions produced, as well as their relative intensity distributions, were similar in both narrow and wideband activation experiments. Thus, it seems that the primary effect of wideband activation is to eliminate the dominance of $[M + H - NH_3]^+$ dissociation products.

The overall results demonstrate that tryptic peptides labeled with amidine groups fragment quite differently from their unmodified counterparts. In both MALDI and electrospray ionization experiments involving singly, doubly, and triply charged amidinated precursor ions, enhanced quantities of y_{n-1} fragment ions are observed. Observation of this dissociation product should prove useful in protein identifications, since the identity of a peptide's N-terminal residue can be used as a database searching constraint.

EXAMPLE 3

Peptide De Novo Sequencing Using a Dual Labeling Strategy

Materials

Hemoglobin (human), α-casein (bovine), and TPCK-treated trypsin (bovine) were obtained from Sigma (St. Louis, MO). Tris-hydroxymethyl)aminomethane (TrizmaBase), S-methylisothiourea hemisulfate, and ammonium hydroxide were also supplied by Sigma (St. Louis, MO). Acetonitrile and trifluoroacetic acid (TFA) were purchased from EM Science (Gibbstown, NJ). Thiopropionamide was supplied by TCI America (Portland, OR). Anhydrous diethyl ether, thioacetamide, and ammonium bicarbonate were purchased from Fisher (Fair Lawn, NJ). Iodomethane, poly (propylene glycol), and formic acid were obtained from Aldrich (Milwaukee, WI). Octadecyl derivatized silica gel (BioBasic 18) was supplied by Thermo Electron (San Jose, CA).

Synthesis of S-methylthioacetimidate.

Thioacetamide (11 g) was dissolved in 1 L of anhydrous diethyl ether. Subsequently, 8.8 mL of iodomethane were added to this solution and the mixture was

allowed to stand at room temperature for 14 h. The precipitate was collected by vacuum filtration and stored over desiccant at ambient temperature without further purification.

Synthesis of S-methylthiopropionimidate

Thiopropionamide (1.8 g) was dissolved in 100 mL of 99.5% pure acetone. This solution was warmed to 60 °C in a water bath before adding 3.8 mL of iodomethane. The reaction mixture was incubated for 1 h, without stirring, at the bath temperature. The product was collected after evaporation of the solvent in a vacuum chamber. The crystals were stored at ambient temperature over desiccant and were not further purified.

Tryptic Digestions

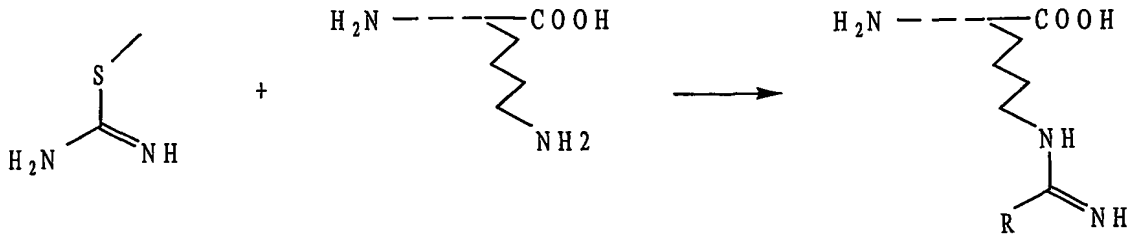
Tryptic peptides from α-casein and hemoglobin were generated using TPCK-treated trypsin. Stock solutions of α-casein and hemoglobin (100 μM) were prepared in 25 mM ammonium bicarbonate. To begin the digestion, 100 μL of protein stock solution were added to 5 μg of lyophilized trypsin and the mixture was stirred. Each digestion was allowed to incubate at 37 °C for 12 h before being stored at -20 °C.

Labeling Reactions

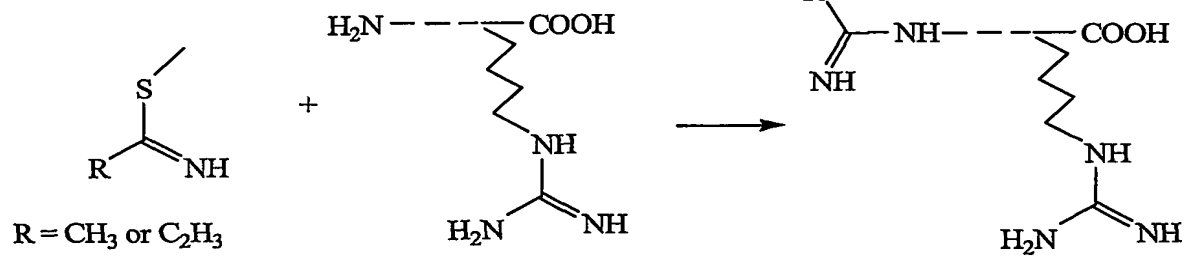
Peptides were derivatized using both guanidination and amidination reactions. First, lysine residues were converted to homoarginines using S-methylisothiourea hemisulfate (Scheme1A). A 1 M mixture of this reagent was prepared in 6% NH_4OH (v/v) and combined with digest solution in a 1:1 ratio (v/v). The reaction mixture was incubated for 1 h at 65 °C. Prior to performing the amidination reactions NH_4OH was removed using a speed-vac (Jouan, Winchester, VA, USA). Next, the guanidinated peptide mixture was reconstituted in H_2O. Acetamidination and propionamidination derivatizations were performed as previously described by Beardsley and Reilly, Journal of Proteome Research 2003, 2, 15-21. (see Scheme 1B). However, only N-termini were labeled since lysine residues were blocked by guanidination. A 43.4 g/L solution of S-methyl thioacetimidate was prepared in 250 mM Trizma Base and mixed 1:1 (v/v) with guanidinated peptides. Similarly, propionamidination reactions were prepared by making 1:1 mixtures of guanidinated peptides and 46.2 g/L S-methyl thiopropionimidate in 250 mM Trizma Base. Each reaction was incubated for 1 h at ambient temperature before acidifying the mixtures by adding TFA to a concentration of 2% (v/v). The reaction mixtures were combined and clean-up was achieved by solid

phase extraction using a 20X1mm BioBasic C_{18} Javelin guard column (Thermo Electron Co., San Jose, CA).

Scheme 1A



Scheme 1B



Liquid Chromatography-Tandem MS of Labeled Peptides

Reversed phase liquid chromatography was performed using a column that was constructed by packing BioBasic C_{18} (Thermo Electron Co., San Jose, CA) media into a 50 mm length of 254 μm ID polyetheretherketone (PEEK) tubing (Upchurch Scientific, Oak Harbor, WA). A linear gradient of increasing acetonitrile was used at a flow rate of 5 μL/min in all LC-MS/MS experiments. This flow rate was established by pre-column splitting of eluent delivered by a Waters 2795 Separations Module (Waters, Milford,MA). Buffer A consisted of 0.1% aqueous formic acid while buffer B was 0.1% formic acid in acetonitrile. All separations were carried out by increasing the concentration of buffer B from 5% to 40% (v/v) over 30 min. The effluent was directed to the electrospray ionization source (Z-spray) of a quadrupole time of flight (Q-TOF) mass spectrometer (Q-Tof micro, Micromass, Manchester, UK). A potential of +3.0 kV was applied to the electrospray needle in all experiments. MS and tandem MS spectra were acquired using the survey scan option provided in the manufacturer's software

(MassLynx). Each scan consisted of spectra that were acquired at a rate of 21.3 kHz and integrated over 1 sec. intervals. The three most intense peaks were selected from each MS scan in real time and subsequently fragmented by low-energy collision induced dissociation (CID). Argon was used as the target gas in all experiments. The number of 1 sec MS/MS scans per precursor ion was limited to five by adding peaks to an exclusion list upon reaching that total. The collision energy (16-50 eV) applied to each precursor ion was varied depending on both charge state and *m/z*.

Results

Fragmentation Properties of Guanidinated/Amidinated Peptides

LC-MS/MS experiments were done using unmodified and doubly labeled tryptic digests of standard proteins to investigate the effects of the derivatizations on peptide fragmentation. The Q-TOF tandem mass spectra of the $[M+2H]^{2+}$ EFTPPVQAAYQK (SEQ ID NO: 3) precursor ion displayed in Figs. 2A &2B are typical examples of this study. Cleavage of the TP peptide bond associated with formation of y_9^{2+}, y_9^{+}, and a series of internal ions was clearly the most efficient fragmentation pathway of the unmodified peptide (Figure 2A). It is well known that peptide bonds adjacent and N-terminal to proline residues are highly labile in CID. Despite the predominance of cleavage between TP a series of y-type ions from y_3 to y_{10} were also observed. Similarly, the spectrum of the labeled peptide (Figure 2B) also displayed these sequence ions. However, the most striking feature of this spectrum is the predominance of complementary b_1 and y_{11} fragment ions resulting from cleavage of the N-terminal peptide bond (EF). These products were not observed from the unmodified peptide whereas they are among the most abundant in the latter example.

Despite the high efficiency of N-terminal peptide bond dissociation the relative intensity distribution of the other sequence ions remains remarkably similar to the unmodified example. Therefore, the enhancement of this dissociation pathway has increased the overall information content of the data, allowing for facile identification of the N-terminal residue using the newly formed b_1 and y_{11} product ions. These data are typical of the fragmentation observed from dozens of peptides that applicants have studied. While the y_{n-1} and b_1 ions are typically not observed from unmodified peptides they are usually the most abundant when the N-terminus is amidinated. In previous work involving CID of amidinated peptides in an ion trap the enhancement of b_1 ions was not discussed. These ions were likely formed in that work, but the low mass cutoff that is

inherent to resonance excitation in an ion trap prevented the analysis of small product ions.

De Novo Sequencing Using Mass Coded N-termini

We present a global *de novo* sequencing strategy that utilizes both acetamidination and propionamidination of peptide N-termini to provide mass signatures that facilitate the discernment of N- and C-terminal fragment ions in MS/MS spectra. Additionally, lysine residues are converted to homoarginines to prevent their amidination. Therefore, C-terminal fragment ions (e.g. y-ions) appear as isobaric pairs in separate MS/MS spectra regardless of the N-terminal label. Since the amidine groups differ by a methylene unit N-terminal fragment ions (e.g. b-ions) are separated by 14 u. This strategy facilitates *de novo* sequencing by eliminating misinterpretations that may be caused by measuring spacings between peaks belonging to different series' of fragment ions. To investigate this approach the tryptic peptides of hemoglobin and α-casein were used. The QTOF tandem mass spectra of derivatized FFVAPFPEVFGK (SEQ ID NO: 3) displayed in Figure 3A & 3B provide a typical example of how the labeling facilitates *de novo* sequencing. The acetamidinated and propionamidinated peptides are represented in Fig. 2A and Fig. 2B respectively. These spectra are remarkably easy to compare since they are nearly identical with regard to the types of fragment ions formed and their intensity distributions. Much like the data of Figure 2A & 2B the complementary b_1/y_{n-1} ions are abundant features in these spectra. CID of this peptide without labeling did not yield either of these fragment ions. These peaks allow the N-terminal residue to be easily identified and provide a valuable starting point for further elucidating sequences. Furthermore, the first two N-terminal residues can easily be interpreted when both y_{n-1} and y_{n-2} are formed. By comparison, the unmodified counterpart of this example also yields y_{n-2}. However, this fragment ion alone does not allow direct interpretation of the first two N-terminal residues. In total, a contiguous y-ion series including y_{11} to y_4 was observed. Due to the mass coding described above the y-ions were easily identified and by using the mass differences between adjacent peaks 66% of this sequence (FFVAPFPE; (SEQ ID NO: 5)) could be determined. In addition to b_1, the b_2* and b_4* (* = neutral loss of NH_3) fragment ions were also observed in both spectra as 14 Da separated pairs. As shown here, with the exception of b_1, other b-type ions are typically accompanied by neutral loss of NH_3. As is common in instruments that employ a collision cell for ion activation some immonium and internal fragment ions were also formed. These ions

include the PEV, PE, and PF products as well as the immonium ions of Phe and Pro. Since these ions are isobaric regardless of the N-terminal label they could be misinterpreted as y-type ions. Fortunately, these peaks are normally isolated to the low mass range and do not interfere with the majority of interpretations.

QTOF tandem mass spectra of the amidinated $[M+2H]^{2+}$ ions of YLGYLEQLLR (SEQ ID NO: 6) were acquired during the analysis of α-casein described above and are displayed in Figs. 4A & 4B. Since this is not a lysine-containing peptide it was not guanidinated during the labeling reactions. However, the N-terminal amino group was amidinated to provide the differential mass signatures. As in the previous example the y_{n-1} (y_9) and b_1 ions are very abundant products formed by the derivatized precursor ions (Figs. 4A and 4B respectively) and the spectra appear qualitatively similar. By matching isobaric peaks in these spectra it was possible to identify the complete y-ion series and therefore infer the entire sequence of this peptide. This peptide also provides an example of how the derivatizations facilitate interpretation of glutamine and lysine residues. The observed mass difference of 128.0558 u between the y_3 and y_4 ions in Fig. 3A is consistent with glutamine. However, since the monoisotopic mass of lysine (128.0950 u) is only 0.0364u heavier than glutamine (128.0586 u) it is not possible to confidently distinguish between the two amino acids unless an instrument with sufficient mass accuracy is used. In the present example there is no ambiguity in the assignment since all lysine residues are expected to have a mass of 170.1168 u following guanidination. The use of guanidination to distinguish these residues would be even more critical if instruments with only unit mass accuracy were used (e.g. ion trap).

CID mass spectra of the $[M+2H]^{2+}$ precursor ions of acetamidinated and propionamidinated LLVVYPW (SEQ ID NO: 7) are displayed in Figs. 5A and 5B respectively. Unlike the examples shown above, this peptide primarily yields b-ions upon CID. This difference in fragmentation behavior is presumably due to the absence of a C-terminal basic residue that can sequester a proton. Much like the previous examples the b_1 ion is a prominent feature in this spectrum. However, the complementary y_{n-1} ion that is typically observed is absent. The y_{n-1} ion is likely formed initially in the same reaction that produces b_1 but undergoes further dissociation to yield the intense y_2 peak resulting from cleavage between YP. The lack of a basic residue such as lysine, arginine or histidine makes it more likely that one of the ionizing protons is located on the peptide backbone, thus facilitating charge-site directed fragmentation of the YP peptide bond. By

using the 14 u mass differentials between peaks in this pair of spectra a b-ion series from b_1 to b_5 was identified. The mass separations between adjacent peaks in this ion series allowed interpretation of LLVVY (SEQ ID NO: 8). The C-terminal residues, PW, were not interpretable from the b-ion series since b_6 was not observed. The suppression of cleavages C-terminal to proline residues is a common attribute in CID and often contributes to incomplete sequence coverage as shown here. However, with slightly more sophisticated data interpretation methods it may be possible to completely sequence peptides from data such as these. In the present case, the b_5 and y_2 ions can be interpreted as a complementary fragment ion pair that is representative of the entire peptide sequence since the sum of their masses is equal to the doubly protonated monoisotopic mass of the precursor ion. Since the absence of fragment ions in CID is most commonly attributable to the presence of proline, a reasonable strategy for interpreting sequence gaps such as these would be to first consider that proline is the next residue. A strategy that combines complementary ion information and consideration of proline will facilitate such interpretations.

LLVVYPW (SEQ ID NO: 7) was produced during tryptic digestion of hemoglobin, but is terminated by tryptophan rather than lysine or arginine. Enzymatic cleavage of the peptide bond C-terminal to aromatic residues is a common side reaction resulting from the chymotryptic specificity of pseudotrypsin that is formed upon tryptic auto-proteolysis. Due to the possibility of non-tryptic peptides, it is often necessary to allow no specificity for cleavage sites when predicting candidate peptides from databases using matching algorithms. The consequences of doing this are that the databases being searched become effectively larger and, as a result, increase the likelihood of false positive assignments. Additionally, data analysis is significantly slower. *De novo* sequencing remains largely unaffected by the presence of non-specific peptides since it involves data interpretation without prior knowledge of database sequences.

<u>De novo sequencing of phosphorylated peptides</u>

Often, one of the primary goals in protein research is to characterize post translational modifications (PTMs). The identification and mapping of these modifications can provide valuable insight into the functional role of proteins. Since PTM sites are not predicted from genomic sequences they are often not identified via database matching algorithms. Therefore, it is important that alternative approaches, such as *de novo* sequencing, be compatible with the study of PTMs. Tryptic peptides from α-

casein were analyzed to test the compatibility of guanidination/amidination labeling with the analysis of phosphorylated peptides. ESI QTOF tandem mass spectra were acquired during an LC separation and Figs 6A & 6B displays the MS/MS spectra of the $[M+2H]^{2+}$ VPQLEIVPN(pS)AEER phosphopeptide (SEQ ID NO: 9). The acetamidinated peptide is shown in Fig. 6A, while Fig. 6B represents the propionamidinated one. In these examples the formation of y-ion minus H_3PO_4 (y_n-ph) was predominant. This effect was also observed in MS/MS spectra of underivatized VPQLEIVPN(pS)AEER (SEQ ID NO: 9)(data not shown) and is a well known artifact of phosphopeptides. Despite the prevalence of H_3PO_4 losses the data are amenable to peptide sequencing because amino acids are identified using the mass spacings in the y-ion series. The observation of a y-ion series from y_{13}-ph to y_7-ph allowed the first seven residues, beginning at the N-terminus, to be sequenced. Furthermore, evidence for the site of phosphorylation was indicated by the appearance of the y_5, y_5-ph, and y_4 ions. The mass difference of 98 u between y_5 and y_5-ph confirms that the y_5 fragment ion contains the phosphorylation site. Also, the observance of y_4, but not y_4-ph, strongly suggests that the C-terminal residue of y_5 was the site of phosphorylation. This residue can be identified because of the 69 u mass differential between y_5-ph and y_4, which corresponds to dephosphorylated serine. This interpretation is consistent with the phosphorylation sites that have previously been reported for this protein. These spectra also demonstrate the deleterious effect that proline residues can have in peptide sequencing. It is well known that dissociation of the C-terminal peptide bond of proline is often suppressed in CID. The low abundance of y_{12}-ph and absence of y_6-ph illustrate this effect. Without observing the y_6-ph ion it was not possible to directly confirm the presence of the proline or asparagine residues in the middle of this sequence. In our experience analyzing tryptic peptides, missed sequence ions are most commonly caused by proline residues. Therefore, a reasonable strategy in *de novo* sequencing may be to consider the presence of proline as a first possibility when interpreting gaps in a series of fragment ions.

Internal Calibration Using b_1 Fragment Ions

As shown above, amidine groups promote fragmentation of the N-terminal peptide bond to produce abundant b_1 and y_{n-1} ions. As demonstrated herein the b_1 ion can serve as an internal calibrant, significantly reducing mass errors in MS/MS spectra. The well known benefits of high mass accuracy in proteomic research apply to all types of

protein identification experiments (e.g. database matching and *de novo* sequencing) since accurate masses allow for tighter constraints, leading to fewer errors.

To examine the effectiveness of this internal calibration approach, LC-MS/MS experiments with guanidinated/amidinated tryptic peptides of simple model proteins were performed using a Q-TOF mass spectrometer. The TOF analyzer was calibrated using PPG prior to the experiment. Following data acquisition MS/MS spectra were internally calibrated using the lock mass utility of the instrument manufacturer's software. This feature calculates the difference between the observed and expected *m/z* of a given ion. The relative error calculated for this peak is then used to correct the calibration of the entire mass spectrum. Therefore, all masses are shifted by an equal percentage of a peak's nominal mass. An example of the effect of this internal calibration method is displayed in Table 1. In this table the mass accuracies of externally and internally calibrated peaks from the CID spectrum of the $[M+2H]^{2+}$ precursor ion of propionamidinated VNVDEVGGEALGR (SEQ ID NO: 10) are compared. Mass errors of about 40 ppm were observed for the peaks of this spectrum prior to internal calibration. In general, these errors were largely dependent on the quality of the external calibration, as well as how recently it was performed. Mass errors commonly drift with increasing time between calibration and analysis due to temperature fluctuations and the instability of power supplies. Regardless of this instability, the errors observed following the correction using b_1 ions were typically less than 10 ppm. The mass accuracies shown in Table 1 demonstrate this improvement.

Table 1: Effect of Mass Correction Using the b_1 ion

VNVDEVGGEALGR		*external calibration only*		*b_1 lock mass*
fragment i.d.	calculated (M+H)	measured (M+H)	mass error (ppm)	mass error (ppm)
y_{12}	1215.5969	1215.5438	39.98	0.82
y_{11}	1101.5540	1101.5155	34.95	-7.90
y_{10}	1002.4855	1002.4476	37.81	-5.09
y_9	887.4586	887.4226	40.57	-2.37
y_8	758.4160	758.3840	42.19	-0.79
y_7	659.3476	659.3187	43.83	1.06
y_6	602.3261	602.3038	37.02	-5.98
b_5*	595.2728	595.2478	42.00	-0.84
b_4*	466.2301	466.2104	42.25	-0.64
y_4	416.2621	416.2477	34.59	-8.17
b_1	155.1184	155.1117	43.19	0.00

Without the use of FT-ICR MS, mass accuracies less than 10 ppm are difficult to routinely achieve unless some form of internal calibration is applied. However, during the course of a typical proteomic experiment it is difficult to implement internal calibrations since they require mixing a calibrant online with LC effluent prior to mass analysis. Not only does this method dilute the effluent, but it is very difficult to match the calibrant and analyte concentrations. Furthermore, including calibrant masses in MS/MS spectra is not feasible since precursor ion isolation necessarily excludes other masses. A quasi-internal calibration alternative to on-line mixing has been to use a dual ESI source in which one source sprays the LC effluent while the other contains a reference compound of known mass (i.e. Lock Spray). In such an experiment the reference channel is intermittingly sampled and mass corrections are made in real-time based on the errors observed for this ion. A drawback of this approach is that the use of a separate reference channel reduces the analyte duty cycle, which is often critical in proteomic investigations involving complex mixtures.

The use of b_1 ions for calibration overcomes the disadvantages of those approaches described above since it is available without online mixing or the introduction of a second ionization source. This ion is well suited to be a calibrant because it is limited to the nineteen unique masses representing the common amino acids. Furthermore, b_1 is easy to identify because it is ubiquitously observed and typically appears as one of the most intense peaks upon CID of amidinated peptides. Therefore, this method should be generally applicable in the analysis of complex peptide mixtures. Also, the high intensity of b_1 ion peaks reduces the effects of isobaric chemical noise that could otherwise distort peak shapes and lead to errors in calculating the centroided masses of calibrants.

Protein Identification

The utility of the presently described *de novo* sequencing strategy to proteomics was considered by using two model proteins (α-casein and hemoglobin) and comparing their interpreted peptide sequences to a large database. These proteins, rather than a complex mixture of unknowns, were employed in this work because their sequences and posttranslational modifications are well characterized. The database matching results are displayed in Table 2.

Table 2: Protein Identification Using Tryptic Digests of Model Proteins

α-casein (Bos taurus)

Actual Sequence[1]	Interpreted Sequence[2]	Sequence Matches	Sequence and Precursor Mass
FFVAPFPEVFGK	FFVAPPFPE	1	α-casein, S1 precursor
YLGYLEQLLR	YLGYLEQLLR	1	α-casein, S1 precursor
FVAPFPEVFGK	FVAPF	13	α-casein, S1 precursor
LYQGPIVLNPWDQVK	GPLVLNP	1	α-casein, S1 precursor
FALPQYLK	FALPQ	4	α-casein, S2 precursor
VPQLEIVPN(pS)AEER	VPQLELV	1	α-casein, S1 precursor
LLYQEPVLGPVR	LLYQEPVL	1	β-casein
HQGLPQEVLNENLLR	HQGLPQEVLNEN	1	α-casein, S1 precursor
EMPFPK	EMPFPK	1	β-casein
hemoglobin (Homo sapiens)			
EFTPPVQAAYQK	EFTPPVQAA	1	hemoglobin β-chain
VNVDEVGGEALGR	VNVDEVGGE	1	hemoglobin β-chain
MFLSFPTTK	MFLSF	15	hemoglobin α-chain
FLASVSTVLTSK	FLASVSTVL	1	hemoglobin α-chain
FFESFGDLSTPDAVMGNPK	GDLSTPDAVM	1	hemoglobin β-chain
VLGAFSDGLAHLDNLK	AHLDNLK	3	hemoglobin β-chain
			delta globin
LLVVYPWTQR	LLVVYPW	12	hemoglobin β-chain
			delta globin epsilon globin G-gamma globin A-gamma globin hemoglobin β (Rattus norvegicus) hemoglobin, β (Bos taurus) hemoglobin, β adult major chain (Mus musculus) hemoglobin, β adult minor chain (Mus musculus) epsilon Y globin (Mus musculus) similar to hemoglobin β-1 chain (B_1)(Major)(Mus musculus) hemoglobin, epsilon 1 (Sus scrofa)
LLVVYPW	LLVVY	29	hemoglobin β-chain delta globin epsilon globin G-gamma globin A-gamma globin hemoglobin beta (Rattus norvegicus) hemoglobin, β (Bos taurus) hemoglobin, β adult major chain (Mus musculus) hemoglobin, β adult minor chain (Mus musculus) epsilon Y globin (Mus musculus) similar to hemoglobin beta-1 chain (B_1)(Major) (Mus musculus) hemoglobin, epsilon 1 (Sus scrofa)

[1] The sequence identifiers for each sequence are as follows:

FFVAPFPEVFGK (SEQ ID NO: 4); YLGYLEQLLR (SEQ ID NO: 6); FVAPFPEVFGK (SEQ ID NO: 11); LYQGPIVLNPWDQVK (SEQ ID NO: 13); FALPQYLK (SEQ ID

NO: 15); VPQLEIVPN(pS)AEER (SEQ ID NO: 9); LLYQEPVLGPVR (SEQ ID NO: 18); HQGLPQEVLNENLLR (SEQ ID NO: 20); EFTPPVQAAYQK (SEQ ID NO: 3); VNVDEVGGEALGR (SEQ ID NO: 10); MFLSFPTTK (SEQ ID NO: 25; FLASVSTVLTSK (SEQ ID NO: 27); FFESFGDLSTPDAVMGNPK (SEQ ID NO: 29); VLGAFSDGLAHLDNLK (SEQ ID NO: 2); LLVVYPWTQR (SEQ ID NO: 32); and LLVVYPW (SEQ ID NO: 7).

[2] The sequence identifiers for each sequence are as follows:

FFVAPFPE; (SEQ ID NO: 5); YLGYLEQLLR (SEQ ID NO: 6); FVAPF (SEQ ID NO: 12); GPLVLNP (SEQ ID NO: 14); FALPQ (SEQ ID NO: 16); VPQLELV (SEQ ID NO: 17); LLYQEPVL (SEQ ID NO: 19); HQGLPQEVLNEN (SEQ ID NO: 21); EMPFPK (SEQ ID NO: 22); EFTPPVQAA (SEQ ID NO: 23); VNVDEVGGE (SEQ ID NO: 24); MFLSF (SEQ ID NO: 26) FLASVSTVL (SEQ ID NO: 28); GDLSTPDAVM (SEQ ID NO: 30); AHLDNLK (SEQ ID NO: 31); LLVVYPW (SEQ ID NO: 7); LLVVY (SEQ ID NO: 8).

For each peptide the interpreted sequence was submitted to a Blast search of the NCBI reference sequence database. In all searches this database was constrained to mammalian proteomes only, which included a total of 81,351 protein sequences. Although Blast searching provides the capability to match homologous sequences, only exact matches were accepted as assignments. Since leucine and isoleucine are isobaric, matching database proteins that contained either residue were treated equally. The number of exactly matching sequences is displayed for each peptide. In most cases (11 of 17) there was sufficient sequence coverage to uniquely match a single protein. However, there were a few examples in which only short segments of their sequences were interpretable, leading to random matches. To resolve this issue of false-positive assignments the precursor mass was also employed as a constraint. Therefore, a sequence match was only considered an assignment if the interpreted sequence was contained within a predicted peptide that was consistent with the observed precursor ion mass. As shown in Table 2 the use of this simple constraint eliminated false positive matches and uniquely identified the model proteins. Since many, nearly identical variants of the β-chain of hemoglobin exist, multiple matches were observed even after consideration of both precursor mass and interpreted sequence. Furthermore, the matches to hemoglobin in other organisms are reported here as well. In a typical experiment it would be possible to eliminate these matches since the organism under study would be known.

Although using the precursor masses and interpreted sequence was sufficient in this work, it may be necessary to further constrain some searches. This would be especially important if only a short segment of a peptide (i.e. < 5 residues) was

interpretable. As demonstrated in Table 2 most interpreted sequences begin with the N-terminal residue. It is clear that these sequences contain the N-terminus since the analysis begins with the b_1 and y_{n-1} fragment ions that are produced by amidinated peptides. In cases such as these, it would be possible to further limit random matches by requiring that the N-terminus of candidate peptides is contained in the interpreted sequence. Another strategy to further refine assignments would be to use smaller fragments of interpretable sequences in addition to the contiguous sequences shown here. In all of the interpretations shown in Table 2 the longest contiguous sequence that was interpretable was matched against a database. However, it is often possible to identify shorter portions of a peptide sequence as well. Incorporation of this additional sequence information could be useful, especially in cases where a long contiguous sequence is not interpretable.